remote sensing

MDPI

Article

SAFDet: A Semi-Anchor-Free Detector for Effective Detection of Oriented Objects in Aerial Images

Zhenyu Fang [1,2], Jinchang Ren [1,2,*], He Sun [2,3], Stephen Marshall [2], Junwei Han [4] and Huimin Zhao [1]

1 School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510665, China; zhenyu.fang@strath.ac.uk (Z.F.); zhaohuimin@gpnu.edu.cn (H.Z.)
2 Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XQ, UK; hesun@bit.edu.cn (H.S.); stephen.marshall@strath.ac.uk (S.M.)
3 School of Automation, Beijing Institute of Technology, Beijing 100081, China
4 School of Automation, Northwestern Polytechnical University, Xi'an 710109, China; jhan@nwpu.edu.cn
* Correspondence: jinchang.ren@strath.ac.uk

Received: 24 August 2020; Accepted: 29 September 2020; Published: 3 October 2020

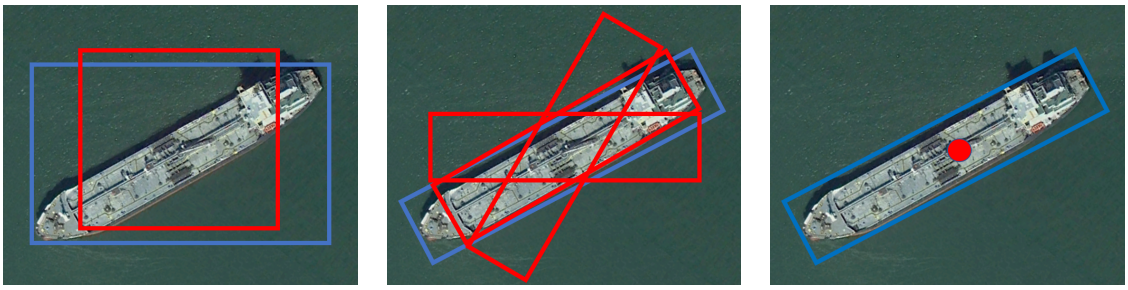
Abstract: An oriented bounding box (OBB) is preferable over a horizontal bounding box (HBB) in accurate object detection. Most of existing works utilize a two-stage detector for locating the HBB and OBB, respectively, which have suffered from the misaligned horizontal proposals and the interference from complex backgrounds. To tackle these issues, region of interest transformer and attention models were proposed, yet they are extremely computationally intensive. To this end, we propose a semi-anchor-free detector (SAFDet) for object detection in aerial images, where a rotation-anchor-free-branch (RAFB) is used to enhance the foreground features via precisely regressing the OBB. Meanwhile, a center-prediction-module (CPM) is introduced for enhancing object localization and suppressing the background noise. Both RAFB and CPM are deployed during training, avoiding increased computational cost of inference. By evaluating on DOTA and HRSC2016 datasets, the efficacy of our approach has been fully validated for a good balance between the accuracy and computational cost.

Keywords: rotate region; convolutional neural network; anchor free; aerial object detection

1. Introduction

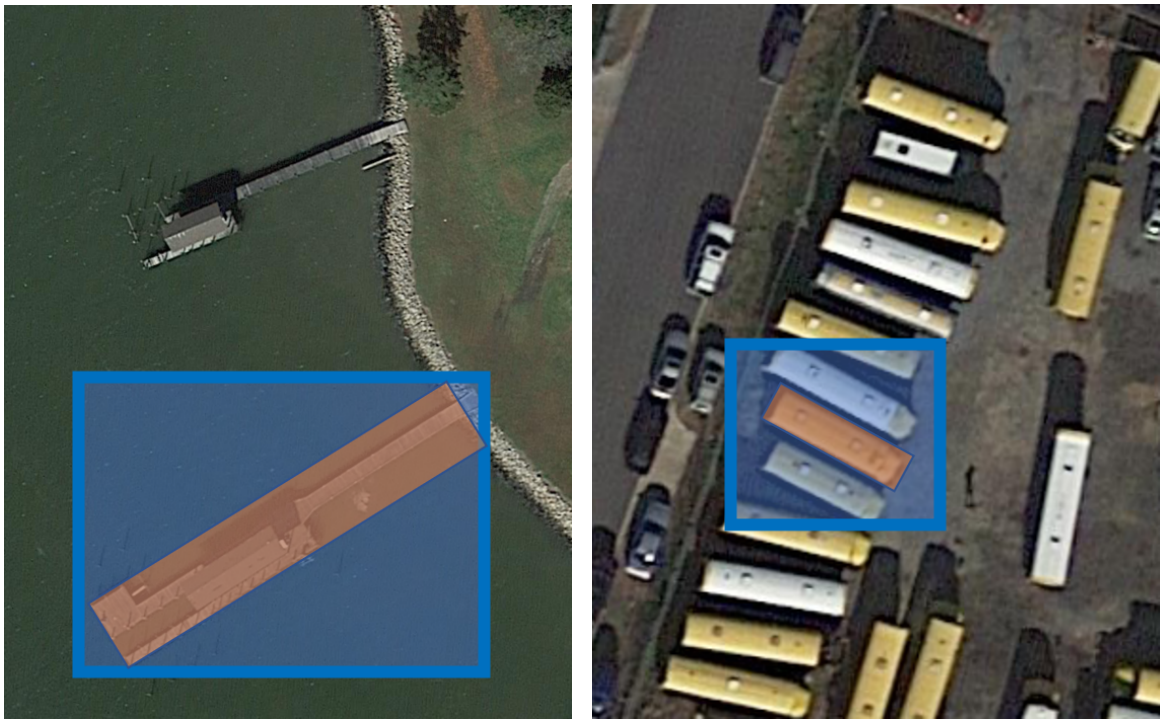
Object detection is one of the basic tasks in computer vision. In aerial images, object detection tends to locate and identify objects from bird's-eye views. Differently from general object detection [1,2], the bounding boxes in aerial images are arbitrarily oriented, rather than horizontally, as in other images. Furthermore, the complex background in aerial images increases the difficulty of feature extraction, which makes the detector have difficulty in distinguishing the (small) foreground objects. Existing methods, such as RICNN [3], R2CNN [4], ROI-Transformer [5] and TextBoxes++ [6], have reached promising results by utilizing two-stage detectors [7–9], i.e., to select the regions of interest (ROIs) via a region proposal network (RPN) and to then use a two-branch subnet for subsequent category classification and box regression.

Most of existing methods predict the ROIs based on horizontal anchors [9]; see Figure 1a. However, as horizontal proposals are located along the image edge, the extracted feature of an object may contain features of the background and the surrounding objects, especially for objects with high aspect ratios or those densely positioned. The misalignment [5,10] between extracted features and object features will interfere with the following detection procedure. Examples of such misalignment are shown in Figure 2, wherein the majority area of the bounding box is the background (a) or the surrounding objects (b).



(**a**) Horizontal anchor (**b**) Rotate anchor (**c**) Proposed method

Figure 1. Examples of ROI prediction using horizontal anchor (**a**), rotation anchor (with multiple rotation angles) (**b**) and the proposed RAFB (**c**), respectively. In (**a**,**b**), only anchors with a single combination of scale and aspect ratio are shown for clarity; i.e., there were originally multiple combinations of scales and aspect ratios in each pixel. Anchors are denoted by red, while the predicted ROIs are labeled by blue. As the rotation ROIs in the proposed RAFB are predicted anchor-free, the predicting pixel is specially highlighted by red. As seen, compared with horizontal anchor-based methods, with the proposed RAFB, the number of anchors in terms of rotation-anchor-based methods is significantly increased. As a result, the total computational cost is increased.



(**a**) (**b**)

Figure 2. Examples of misalignment caused by (**a**) a background feature and (**b**) features from surrounding objects. In each example, the extracted feature(s) is highlighted in blue and the object feature is colored in orange.

To tackle the misalignment problem, existing methods usually have two solutions; i.e., either replacing the horizontal anchors with the rotation anchors, e.g., Liu et al. [11], RRPN [12] and R2PN [13] (as shown in Figure 1b), or enhancing the capability of noise suppression [5,14,15]. However, on the one hand, rotation anchors need to take the predefined angle (or orientation of object) into account; hence, the anchor number can be dramatically increased. On the other hand, the capability of feature extraction relies on certain assistant modules [14,16,17], such as multiple convolutional

layers for enhanced feature extraction. As a result, both solutions mentioned above increase the total computational cost and are not ideal in practical applications due to the poor efficiency.

In summary, horizontal anchor-based detectors are useful but suffer from misalignment-caused inaccuracy. Existing works tackle this issue by either utilizing rotated anchors or implementing additional feature extraction modules. However, both of these solutions will increase the computational cost, leading to degraded efficiency. As a result, we aim to alleviate the misalignment caused by horizontal anchors without sacrificing the computational efficiency.

In this paper, we propose a semi-anchor-free detector (SAFDet) to tackle the aforementioned challenges via multi-task learning. In the proposed SAFDet, RPN learns rotational ROI (RROI) prediction and feature enhancement through two proposed modules: a rotation-anchor-free branch (RAFB) and a center prediction module (CPM). To reduce the computational cost caused by rotated anchor, RAFB is utilized, as illustrated in Figure 1c, which directly predicts the oriented bounding box (OBB) without any predefined rotation of the anchor. For suppressing the noise caused by the background, a CPM is proposed, which focuses on the prediction of centroid of the ROI. Experimental results have validated that RAFB improves the detection performance (mAP) by 1.6%. After implementing CPM, mAP is further improved by 1%. As both the proposed modules are only utilized in training, the computational cost of inference is not increased.

The main contributions of this paper can be summarized as follows:

1. A semi-anchor-free detector (SAFDet) is proposed for effective detection of oriented objects in aerial images;
2. A rotation-anchor-free branch (RAFB) is proposed to tackle the misalignment problem when using horizontal anchors, which can predict the OBB directly without any predefined anchor setting;
3. A center prediction module (CPM) is implemented to enhance the capability of feature extraction during the training of RPN, hence avoiding increased computational cost of inference;
4. In experiments on the DOTA [18] and HRSDC2016 [11] datasets, the two-stage detector implemented from the proposed approach achieved superior performance over a number of state-of-the-art methods and reached a good balance between the accuracy and speed.

The remaining parts of this paper are organized as follows. Section 2 briefly introduces the related work. The design of the proposed approach, especially of the RAFB and the CPM modules, is detailed in Section 3. Experimental results, including the ablation study and discussion, are presented in Section 4. Finally, some concluding remarks are made in Section 5.

2. Related Works

2.1. Anchor-Based Oriented Aerial Object Detection

An anchor is similar to the "sliding window" for non-CNN methods [19]. For general object detection, existing methods such as SSD [20], faster-RCNN [9], feature pyramid network [21] and RetinaNet [22], predict the bounding box of the object by regressing the difference between a number of candidate boxes at each pixel location (known as the "anchor") and the actual one (the ground truth) in the image. For object detection in aerial images, however, there is no angle element in the horizontal anchor setting and the rotational angle is fixed to 90 degrees. This increases the difficulty of rotated angle regression. Moreover, horizontal anchors usually contain more background area than rotational anchors, which has resulted in the problem of misalignment between the detected ROI and the ground truth. To address this challenge, rotated anchor setting was introduced in RRPN [12], ICN [15], R2PN [13] and R3Det [17], where anchors at different angles are predefined. However, the increased number of anchors ("$numscales \times numaspectratios \times numangles$" compared with "$numscales \times numaspectratios$") makes it more time-consuming than the horizontal anchor setting in object detection.

2.2. Anchor-Free Object Detection

The anchor-free strategy was originally designed for general object detection [23–26]. Differently from anchor-based methods, such as faster-RCNN [9], SSD [20], FPN [21] and RetinaNet [22], anchor-free methods predict one bounding box per pixel. Instead of regressing the difference between the actual bounding box and the anchor, anchor-free methods predict the bounding box based on the coordinates of the feature map. However, the regression formats are different between anchor-free methods. CornerNet [24] predicts the upper-left and button-right coordinates of the bounding box, known as the "XYXY" mode. YOLO [23] and CSP [27] predict the center coordinate, and the associated width and height of the bounding box, known as the "XYWH" mode, which is the same mode as used in the anchor-based methods. Differently from these two prediction modes, FCOS [26] predicts the distances between the center point and the four boundary lines of the bounding box, respectively. Anchor-free methods can actually reduce the computational cost on the box regression branch. However, due to the lack of anchors in anchor-free methods, the widely used intersection over union (IoU) matching scheme can no longer be applied. For anchor-based methods, proposals are categorized to foreground and background via the IoU value; for anchor-free methods, the foreground and background proposals are empirically defined. Taking YOLO [23] as an example, the center pixels (known as "gird" in YOLO) of the bounding box are defined as foreground, while all others are defined as negative, i.e., background. Such a scheme actually rejects lots of instances, even with high prediction scores, resulting in much lower recall values. Hence, SSD outperforms YOLO [20]. This also indicates that the positive assignment becomes a challenging task for anchor-free methods [26].

2.3. Multi-Task Loss of Oriented Aerial Object Detection

Previous works applied multi-task prediction [10,14,28] to enhance the capacity of box prediction. Apart from bounding box regression and category classification, multi-task prediction can also be applied in other important tasks. In MTCNN [28], the keypoint prediction is utilized to assist face detection. In mask-RCNN [10], an image segmentation branch is used to predict foreground pixels jointly with box regression. In SCRDet [14], attention detection is implemented for suppressing the background noise, which is also utilized in face detection [16]. As mentioned in [18], existing challenges include low resolution, noise and crowding. Previous works tackled these challenges by using attention structures [14,16]; supervised attention structures [14,16] consider all the pixels within the boxes as the foreground. However, the anchors centered at the boundary of the bounding boxes, especially for those with large aspect ratios (height/width), may have lower IoU matching values than the predefined threshold. These boundary pixels are supposed to be categorized as the background rather that the foreground.

For aerial object detection, there are challenging scenarios, including small objects, dense arrangements and freely oriented objects. Thus, in this paper, we present a two-stage aerial object detector based on the semi-anchor-free model. To gain the benefits from both anchor-based and anchor-free regression schemes, we introduce an anchor-free-branch in the region proposal network (RPN) during training and apply anchor-based regression branch during the inference stage. Though we adopted a box based anchor-free regression scheme similar to CSP [27], the center coordinates are shared with the anchor-based prediction, which helps to reduce the regression difficulty; i.e., only the shape regression of bounding box is required instead of both shape and position, in anchor-free prediction. Meanwhile, the proposed anchor-free regression schemes are for prediction of oriented bounding boxes. Thus, they also include the rotation angle regression module, which is not considered by CSP. In order to locate the positions of objects more precisely, we present a center prediction module, which predicts the center pixels of the objects only as detailed in the next section. A similar module is proposed in SCRDet [14]. Differently from the SCRDet, the proposed CPM predicts the center pixels of the bounding boxes, rather than all pixels within the bounding boxes. Furthermore, the proposed CPM is only implemented for training. Thus, it will not increase the computational cost of inference as in SCRDet.

3. The Proposed Method

3.1. Overall Architecture

The flowchart of the proposed SAFDet detector is sketched in Figure 3. Following the work in [5,14], we take the ResNet-v1 [29] as the backbone and denote the outputs for "conv2," "conv3," "conv4" and "conv5" of ResNet as C2, C3, C4 and C5, respectively. Existing two-stage networks [4,5] adopt C4 as the feature map, where the total stride(s) is large (s = 16). As a result, the resolution of the feature map is too coarse for detecting small objects, such as vehicles, ships and bridges. When applying the feature pyramid network (FPN) [21] on the backbone, the total stride will be reduced to 4, causing an increase on the computational cost of the prediction. To balance the accuracy and the computational cost, we use the {C3, C4} in the FPN. Hence, the total stride becomes 8. In the FPN, we only apply a single convolutional layer on C3 to adjust the number of channels without using any other enhancement modules [14,15].

For regressing the bounding box, RPN predicts both horizontal ROIs (HROIs) and rotational ROIs (RROIs). For the prediction of HROIs, we use the horizontal anchor to save computational time. The corresponding RROIs are accomplished by the proposed rotation-anchor-free branch (RAFB), which is highlighted by green in Figure 3. Meanwhile, the proposed center prediction module (CPM) predicts the center coordinates of ROIs from the same feature map used by RPN (denoted by gold in Figure 3). As both RAFB and CPM are utilized during training, only the HROIs will be used in the following ROI pooling layer [9].

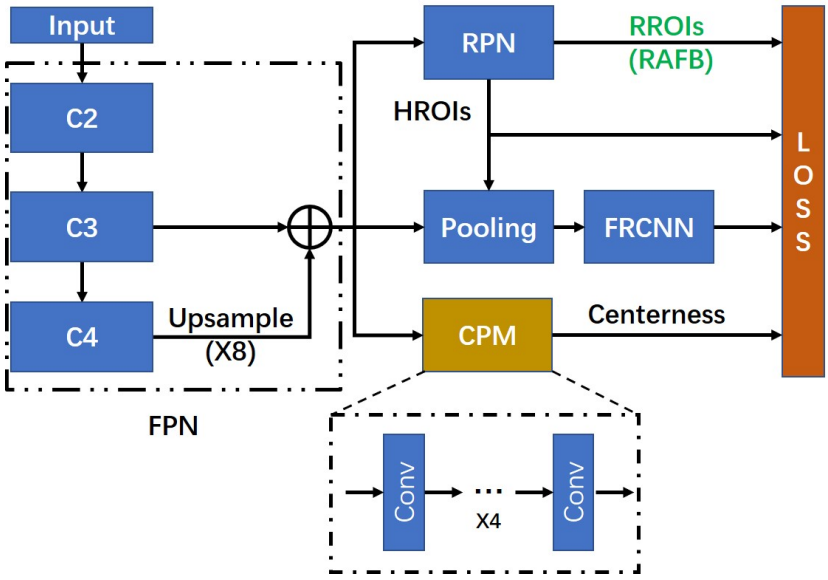


Figure 3. The proposed network structure. The backbone feature pyramid network is denoted by "FPN." The "HROIs" and "RROIs" respectively indicate the horizontal ROIs and rotational ROIs predicted by the region proposal network (RPN). Specifically, RROIs are predicted by the proposed rotation-anchor-free branch (RAFB) in the RPN, which is highlighted by green. Meanwhile, the proposed center prediction module (CPM) predicts the center coordinates of bounding boxes, which is denoted by gold in the figure. As both RAFB and CPM are utilized during training, only the HROIs is used on the pooling layer and the fast RCNN module (FRCNN).

For the fast RCNN (FRCNN), the network predicts both OBB and the corresponding horizontal bounding box (HBB) as in [4,14,17]. Relevant details are presented in the following sections.

3.2. Rotational Anchor-Free Branch

As described in [17], the rotational anchor setting in the RPN achieves a better performance when compared with horizontal box prediction. However, rotate anchor setting generates more boxes

at each point of the feature map, which significantly increases the computational cost at the same time. To improve the performance without increasing the computational cost, for each horizontal box prediction, we additionally predict its corresponding OBB using an anchor-free method, which acts as an auxiliary loss during the training. For the original horizontal branch, we use the typical four-element presentation method [4,9,20]. The four elements can be mathematically expressed as x_c, y_c, w_{hbb} and h_{hbb}, denoting respectively the center coordinates, the width and the height of a box. Similarly, we encode the OBB in the same way as the HBB with one additional element for the rotation angle. As a result, the five elements are $(x_c, y_c, w_{rot}, h_{rot}, \theta)$, where w_{rot} and h_{rot} denote the width and the height of a rotate box, respectively. Consistently with OpenCV [30], θ is the angle acute to the x-axis, ranging in $[-90, 0)$. As the center coordinates of HBB and OBB are the same [18], during training, the prediction of the center coordinates is shared between HBB and OBB. To summarize, the output of the RPN for each box prediction, after the decoding method, has in total seven elements, i.e., x_c, y_c, w_{hbb}, h_{hbb}, w_{rot}, h_{rot} and θ, which is visualized in Figure 4.

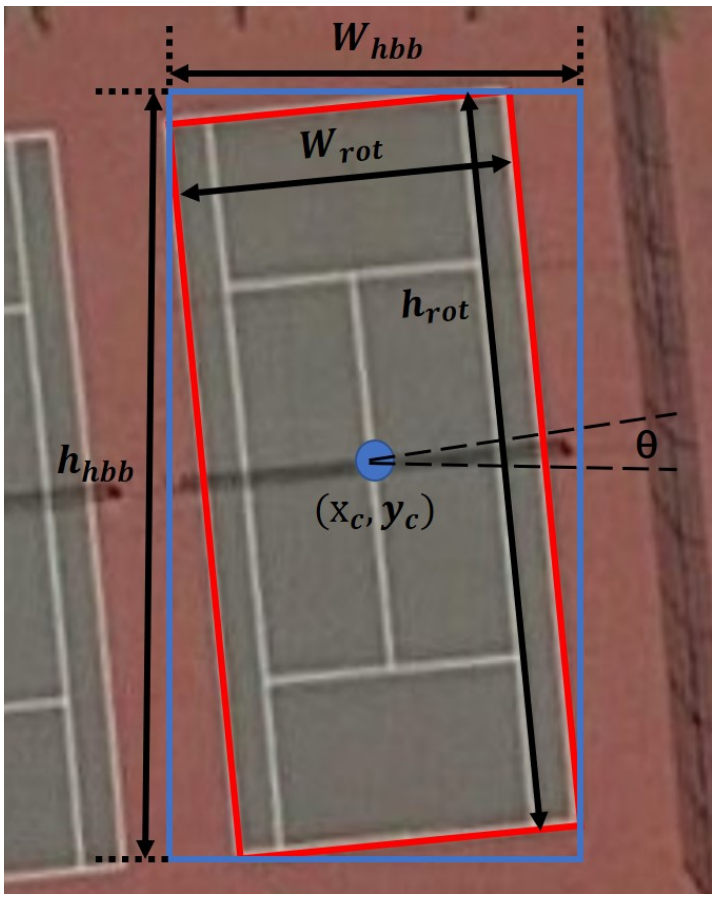


Figure 4. Illustration of the elements predicted by RPN. The horizontal bounding box and the oriented bounding box are highlighted in blue and red, respectively. The horizontal bounding box is represented as x_c, y_c, w_{hbb} and h_{hbb}, denoting respectively the center coordinates, the width and the height of the box. Meanwhile, w_{rot}, h_{rot} and θ denote the width, height and rotational angle of the rotated box, respectively. As the center coordinates of HBB and OBB are the same, the center coordinates will be shared in between.

Regression Scheme of the Rotation Branch

As the rotation branch is anchor-free, we only define horizontal anchors for the RPN. An OBB is assigned as the foreground if its corresponding HBB is assigned as the foreground; i.e., the OBB matching is based on the intersection over union (IoU) value between the horizontal ground truth box and the horizontal anchor, as in [8,20]. For rotational box regression in RPN, we keep the regression

method of the center coordinates from HBB due to coordinate sharing, and directly predict the size (w_{rot}, h_{rot}) and the rotation angle (θ) using:

$$\begin{aligned} w_{rot} &= exp\left(w_{out}\right) * s; \\ h_{rot} &= exp\left(h_{out}\right) * s; \\ \theta &= -90^{\circ} + \frac{\theta_{out}}{\pi} * 180^{\circ}; \end{aligned} \tag{1}$$

where w_{rot}, h_{rot} and θ are the predicted width, height and the rotation angle by detector, respectively; w_{out}, h_{out} and θ_{out} are the corresponding encoded outputs from the prediction layer, respectively; s indicates the total stride of the feature map. Similar to [9,26], we use the exponential function, denoted as "$exp()$" above, to keep the output in positive. As the prediction is conducted on the feature map, which is down-sampled by the total stride (s), the s is multiplied to scale the predicted box size to its actual size. The range of θ is in $[-90, 0)$, the magnitude of which is quite larger than other elements. To balance the magnitude between different elements, the proposed method predicts the radian of the rotation angle and also converts the unit to degree using Equation (1). As a result, the range of θ_{out} is in $[0, \frac{\pi}{2})$.

3.3. Center Prediction Module

As mentioned in [14,18], existing challenges include low resolution, noise, crowding and freely oriented objects. Previous works tackled these challenges by using supervised attention structures [14,16], which consider all the pixels within the boxes as the foreground. However, on the one hand, the anchors where the centers are located at the boundary of the bounding boxes may have lower IoU matching values than the predefined threshold, especially for those boxes with large aspect ratios (height/width). On the other hand, the receptive field of boundary pixels is the same as the center pixel. That means that the boundary pixels contain more background information, which may interfere with the feature extraction. As a result, these boundary pixels are supposed to be categorized as the background rather than the foreground, which indicates that previous labeling methods of attention structures may introduce more noise. To remove the ambiguous instance and enhance the foreground features against the complex background, we propose a supervised center prediction module (CPM), as illustrated in Figure 3. To be more specific, the structure consists of four convolutional layers for feature extraction and one binary prediction layer, which is the same as in RetinaNet [22]. Differently from the existing labeling methods [14,16], only the pixels at the center region are taken as the foreground, and the remaining pixels in the bounding box are all labeled as the background.

As shown in Figure 5, the size of the foreground pixels is defined by $\frac{1}{s}$ of the horizontal bounding box, which can be given by:

$$\begin{aligned} w_{cpm} &= w_{hbb}/s; \\ h_{cpm} &= h_{hbb}/s; \end{aligned} \tag{2}$$

where w_{cpm} and w_{cpm} are the width and the height of the foreground region, respectively. As suggested in [25], such settings do the best job of balancing foreground instances and background instances.

Let the prediction score and the ground truth at the position (i, j) be p_{ij} and y_{ij}, respectively. We deploy the variant focal loss, as in [25]:

$$L_{CPM} = -\frac{1}{N} \sum_{i=1}^{H/s} \sum_{j=1}^{W/s} \alpha_{ij} FL\left(p_{ij}, y_{ij}\right); \tag{3}$$

where H and W denote the height and the width of the original image, respectively; s is the total stride, which is the same as in Equation (1); N is the normalization factor. Specifically, $FL(*)$ is the original focal loss [22] and α_{ij} is the adjusting factor:

$$FL\left(p_{ij}, y_{ij}\right) = \begin{cases} \left(1-p_{ij}\right)^2 \log\left(p_{ij}\right), \text{ if } y_{ij}=1 \\ \left(p_{ij}\right)^2 \log\left(1-p_{ij}\right), \text{ otherwise} \end{cases}; \tag{4}$$

$$\alpha_{ij} = \begin{cases} 1, \text{ if } y_{ij}=1 \\ \left(1-\alpha_G\right)^4, \text{ otherwise} \end{cases}; \tag{5}$$

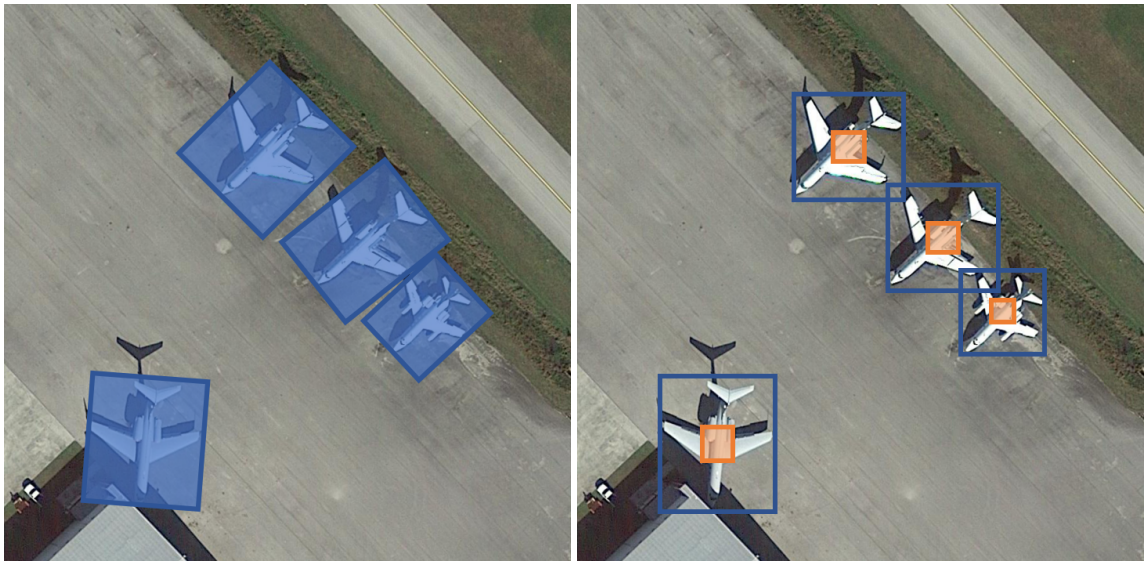
$$\alpha_G = 1 - \max_{k=1,2,\ldots,K} G_k\left(i, j, x_c^k, y_c^k, w_k, h_k\right); \tag{6}$$

where K is the total number of objects. (x_c^k, y_c^k), w_k and h_k are the center coordinates, width and height of the horizontal bounding box k, respectively. $G_k(*)$ is a 2D Gaussian mask centered at the bounding box k, which is formulated as:

$$G_k\left(i, j, x_c^k, y_c^k\right) = \exp\left(-\left(\frac{\left(i-x_c^k\right)^2}{2\sigma_{w_k}^2} + \frac{\left(j-y_c^k\right)^2}{2\sigma_{h_k}^2}\right)\right); \tag{7}$$

where the variances $(\sigma_{w_k}^2, \sigma_{h_K}^2)$ are proportional to the widths (w_k) and heights (h_k) of individual objects.

Experimentally, assigning N as the number of all instances on the feature map gives a better performance than the original setting, where the loss is normalized by the number of objects in the image.



(**a**) (**b**)

Figure 5. The labeling methods of (**a**) the previous attention methods and (**b**) the proposed CPM. The foreground pixels are denoted by blue and orange, respectively. For the proposed CPM, the range of the foreground pixels is defined by the $\frac{1}{s}$ of the horizontal bounding box.

3.4. Loss Function

In summary, the total loss of the network is defined as:

$$L_{tot} = \lambda_1 L_{CPM} + \lambda_2 L_{rotaf} + L_{RPN} + L_{FRCNN} \tag{8}$$

where λ_1 and λ_2 are the weight factors; L_{CPM} and L_{rotaf} are the losses of center prediction module and the anchor-free rotate box prediction branch, respectively. L_{RPN} is the total loss of the region proposal

network as in [9]. L_{FRCNN} is the total loss of the fast RCNN subnet [9], including HBB regression loss, OBB regression loss and the category prediction loss as in [4,14,17]. Experimental results indicate that the magnitude of L_{CPM} is 0.01, the same as the L_{RPN} and L_{FRCNN}, while the magnitude of L_{rotaf} is 0.1. As suggested by SCRDet [14] and SSD [20], to achieve the best performance, the magnitudes of loss elements are supposed to be the same. Thus, to balance the magnitudes between the loss elements, we used $\lambda_1 = 1$ and $\lambda_2 = 0.1$ during the following experiments.

4. Experiments and Analysis

Two benchmark datasets, DOTA [18] and HRSC2016 [11], were chosen for ablation study and performance evaluation, which are well known for oriented object detection. Relevant details are presented as follows.

4.1. Datasets and Implementation Details

4.1.1. DOTA

The benchmark DOTA (dataset for object detection in aerial images) [18] is an aerial image dataset for object detection. Images of DOTA were sampled from multiple sensors and platforms, where the sizes range from around 800 × 800 to 4000 × 4000 pixels. There exist 2806 aerial images and 188,282 instances in total. The annotations of objects concern various scales, oriented angles and aspect ratios, which are classified in 15 commonly used categories. The names of these categories and their abbreviations used in the paper are PL—plane; BD—baseball diamond; BR—bridge; GTF—ground field track; SV—small vehicle; LV—large vehicle; SH—ship; TC—tennis court; BC—basketball court; ST—storage tank; SBF—soccer—ball field; RA—roundabout; HA—Harbor; SP—swimming pool; and HC-Helicopter. The ratios of the training set, validation set and test set are 1/2, 1/6 and 1/3 of the total number of samples, respectively. For the instances in DOTA, about 98% of them are less than 300 pixels, which is similar to real-world scenes. Meanwhile, the sizes of boxes between in different categories—e.g., 50 for a small-vehicle, and 1200 for a bridge. Such a huge variation in terms of box size has inevitably increased the difficulty of accurate object detection. There were two detection tasks done on the DOTA dataset; task 1 was for OBB regression, and task 2 was for HBB regression. As the aim of this study was improving the performance of OBB detection, our experiments were focused on task 1.

4.1.2. HRSC2016

The HRSC2016 dataset (high resolution ship collections 2016) [11] collected 1061 images with 2976 annotated instances from six famous harbors, which are widely used for ship detection. The scenarios contain ships both on the sea and close inshore. In the HRSC2016 dataset, there are over 25 types of ships with large varieties in scale, shape and rotational angle. The image size ranges from 300 × 300 to 1500 × 900 pixels. After excluding hovercrafts and submarines, the sizes of the training set, validation set and test set were 436 images with 1207 instances, 181 images with 541 instances and 444 images with 1071 instances, respectively.

4.1.3. Implementation Details and Evaluation Metrics

When training on the DOTA dataset, the original images were uniformly cropped to patches of 800 × 800 pixels with a stride of 200 pixels. In our ablation study, only the training set was utilized, and the result was evaluated on the validation set. When compared with other detectors, both the training and validation sets were applied for training, and the results were tested on the test set, which were then submitted to the official evaluation server for evaluation. For image pre-processing, each images patch was first subtracted by the mean values of the ImageNet on the RGB channels. Afterwards, randomly flipping was adopted with a probability of 0.5. The learning rate was initialized

as 3×10^{-4} and was divided by 10 at 100,000 and 200,000 iterations, respectively. The training terminated at 300,000 iterations.

For the HRSC2016 dataset, both the training and validation sets were utilized for training. The training images were resized to $(800, 1024)$, where 800 and 1024 indicate, respectively, the minimum and the maximum sizes for the images. The pre-processing steps were the same as used for training on DOTA. Thus, no additional augmentation methods, e.g., image pyramid and image rotation, were applied. The total iterations numbered 12,000. The initial learning rate was 5×10^{-4}, which was divided by 10 over 9000 iterations. Results were tested on the original image size, and then evaluated using the standard VOC-style AP metrics [5,17].

4.2. Ablation Study

A series of ablation studies were conducted to validate the effect of each component in the proposed approach. During the ablation study, the ResNet-50 was taken as the backbone, and relevant models were implemented using Tensorflow [31].

4.2.1. Baseline Setup

As the proposed method is an improved two-stage object detector, we built the baseline starting from the faster-RCNN. Previous works [4,5,14] indicate that jointly predicting the OBB and its boundary rectangle (HBB) can improve the performance of OBB prediction. As a result, in the baseline, HBB is also predicted during training, which is actually an example of R2CNN [4] with a single ROI pooling kernel rather than multiple ROI pooling kernels in the original R2CNN.

The total stride size is another key factor that affects the performances of object detectors. Faster-RCNN sets the total stride size as 16, which is too large for detecting small objects (e.g., ships, small vehicles and bridges) for DOTA. Thus, a top-down pathway is essential for increasing the resolution of the feature map [14,21]. With a decreased total stride size, the accuracy increases, especially for small object detection, as highlighted in Table 1. As can be seen, the R2CNN has the lowest computational speed when the total stride size is 16, though it achieves the highest mean average precision (mAP) when the total stride size is 4. As a result, to balance the computational cost and the accuracy, the total stride size was set to 8 in the baseline R2CNN.

Table 1. The performances of R2CNN with different total stride sizes. Models were trained and validated on the DOTA dataset (%). Computational time were measured on a single V100 GPU with the input size 800×800.

Total Strides	mAP	Time (ms)	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
16	56.22	**65**	86.09	65.02	22.73	49.46	38.26	49.84	52.26	90.44	53.60	74.03	61.95	61.70	55.14	47.92	34.88
8	61.53	72	88.17	68.54	30.64	53.10	49.41	59.43	63.76	90.73	52.34	83.29	63.80	62.09	63.24	52.71	41.83
4	**65.08**	238	88.99	71.13	40.54	61.83	51.60	66.99	68.33	90.73	59.81	83.65	63.65	62.72	64.60	52.92	48.72

4.2.2. Effect of Rotation on the Anchor-Free Branch

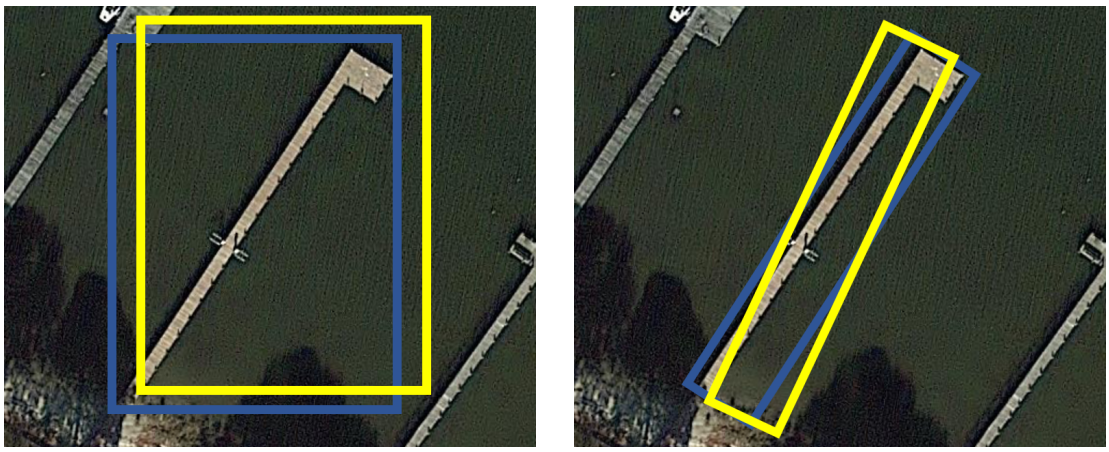
As discussed in Section 3.2, multi-task bounding box regression brings about performance improvements. The effect of rotation-anchor-free-branch (RAFB) is summarized in Table 2. As seen, the performance is improved by 1.59% without increasing the computational cost of the detector. As for each category, RAFB further improves the detection performance in instances with large aspect ratios and/or low resolutions, e.g., 5.7% for bridges, 5.8% for helicopters and 7.5% for ground field tracks.

Table 2. The effect of each component in the proposed method on the DOTA dataset (%), where "RAFB" indicates the rotational anchor-free branch; "Att" is the attention module and "CPM" denotes the center prediction module. "Loss only" means that the attention module is only applied during training, as an auxiliary loss.

Methods	mAP	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
R2CNN-8 (Baseline)	61.53	88.17	68.54	30.64	53.10	49.41	59.43	63.76	90.73	52.34	83.29	63.80	62.09	63.24	52.71	41.83
+RAFB	63.12	88.59	68.57	36.92	61.68	48.65	61.57	57.86	90.53	60.44	79.30	71.95	63.75	58.47	50.88	47.67
+RAFB+ATT	63.33	88.81	72.28	38.04	64.10	47.33	60.37	58.67	90.65	57.89	79.01	72.01	64.29	57.48	51.54	47.53
+RAFB+ATT (loss only)	62.46	88.86	69.86	35.01	62.87	48.59	60.49	58.91	90.32	55.10	79.27	69.34	60.56	57.96	51.80	47.35
+RAFB+CPM	**64.17**	88.63	71.34	37.43	63.51	47.11	62.04	58.62	90.71	64.20	79.23	74.91	63.46	58.06	52.86	50.44

Anchor vs. Anchor-Free

We also trained the network using RAFB for prediction on RPN, which directly predicts the OBB rather than the HBB. As the original IoU matching scheme is unsuitable for anchor-free methods, we simply used the matching scheme as in CSP. However, the mAP was only 38%. This is different from the observations on HBB detection [25,26], which showed that anchor-free methods outperform anchor-based methods. We deduce that was caused by two things. First, for two-stage HBB detection, a slight shift in the x–y coordinates does not significantly affect the feature extraction, as most area of object is still within the bounding box. However, for OBB detection, a slight error in the rotational angle will cause the missing of most of the regions, leading to information loss on feature extraction, especially for instances with large aspect ratios. The information loss on feature extraction will then affect the performance of the second stage prediction. Similar observations on the effect of angle error are reported in [4,5]. An example of HBB detection and OBB detection on the same object is shown in Figure 6. Second, the CSP matching scheme was originally designed for HBB box prediction. As proposed in [26], the matching scheme is vitally important for anchor-free detectors, which motivates the future work for proposing an anchor-free matching scheme for OBB detection.



(**a**) HBB (**b**) OBB

Figure 6. Example of bounding box prediction on HBB (**a**) and OBB (**b**). Ground truth bounding boxes are denoted by blue, while predicted bounding boxes are colored by yellow. Only the centre target are labelled for clarity.

4.2.3. The Effect of the Center Prediction Module

As previously discussed, the center prediction module (CPM) is beneficial to locating the centers of objects. The CPM is only implemented during training; hence it will not increase the computational cost of inference, as shown in Table 2. With the CPM, the mAP of detector is further improved by 1%. Compared with RAFB, which improves the performance on low-resolution objects, CPM can further

bring benefits to the detection of large-scale objects, e.g., baseball diamonds, basketball courts and soccer fields.

An additional experiment was conducted, in which CPM was replaced by the attention module, and the results are shown in Table 2 for comparison. When adopting the attention module into the detector, the mAP increases by 0.2% only. On the one hand, this improvement is not as significant as when CPM is applied. On the other hand, it also increases the computational cost of the detector. When assigning the attention module as an auxiliary loss, the mAP drops by about 0.6%. Therefore, the results indicate that CPM is more effective than the attention module for OBB detection in aerial images.

4.3. Comparison with the State-of-the-Art Detectors

4.3.1. DOTA

When compared with the state-of-the-art methods, for a fair comparison, multi-scale image detection is popularly utilized [14,17], where the short side of the image is resized to 600, 700, 800, 900, 1000, 1100 and 1200, respectively. A deeper backbone, that of ResNet-101, was adopted for a fair comparison. We denote the proposed semi-anchor-free detector "SAFDet." As shown in Table 3, The feature enhancement modules proposed in [5,14,17] improve accuracy but cost through the significantly decreased computational speed. SAFDet lags SCRDet by about 1.3%, but it has no additional feature extraction or fusion module, reaching a good balance between the accuracy and computational cost. Experimental results in [17] indicate that SCRDet improves the accuracy at a cost of much-degraded efficiency, which is also validated in Table 4. As a result, the speed of SCRDet is lower than that of R3CNN. On the contrary, a comparison of inference speed, as detailed in the next section, shows that the proposed method is much faster than R3Det.

Specifically, we compared the proposed SAFDet with several of the most recent two-stage methods. Though RADet [32] achieved state-of-the-art accuracy for the category of swimming pool, the performance when detecting simple-texture objects was not convincing (plane, baseball diamond and soccer field). As a result, the proposed SAFDet outperformed it by 2.3% on mAP. FFA [33] outperformed SAFDet by 0.5%; this is because FFA utilizes an additional feature pyramid pathway, resulting in extra computational cost. When using the same feature pyramid structure, FFA reached 67.90% in mAP without using image pyramid, while SAFDet achieved 70.10%. This further validates that SAFDet attains a good balance between computational cost and inference speed.

Table 3. Quantitative comparison with the state-of-the-art methods on the task 1 of DOTA (%).

Methods	FR-O [18]	R2CNN [4]	RRPN [12]	ICN [15]	RADet [32]	RoI-Trans. [5]	FFA [33]	SCRDet [14]	R3Det [17]	SAFDet (Proposed)
mAP	52.93	60.67	61.01	68.20	69.09	69.56	71.80	72.61	71.69	71.33
PL	79.09	80.94	88.52	81.40	79.45	88.64	89.60	**89.98**	89.54	89.78
BD	69.12	65.67	71.20	74.30	76.99	78.52	76.40	80.65	**81.99**	80.77
BR	17.17	35.34	31.66	47.70	48.05	43.44	48.20	**52.09**	48.46	46.61
GTF	63.49	67.44	59.30	70.30	65.83	75.92	58.90	68.36	62.52	**75.31**
SV	34.20	59.92	51.85	64.90	65.46	68.81	67.20	68.36	**70.48**	66.47
LV	37.16	50.91	56.19	67.80	74.40	73.68	**76.50**	60.32	74.29	60.41
SH	36.20	55.81	57.25	70.00	68.86	83.59	**81.40**	72.41	77.54	66.43
TC	89.19	90.67	90.81	90.80	89.70	90.74	90.10	**90.85**	90.80	90.27
BC	69.60	66.92	72.84	79.10	78.14	77.27	83.30	**87.94**	81.39	82.94
ST	58.96	72.39	67.38	78.20	74.97	81.46	83.40	**86.86**	83.54	85.88
SBF	49.4	55.06	56.69	53.60	49.92	58.39	55.70	**65.02**	61.97	63.19
RA	52.52	52.23	52.84	62.90	64.63	53.54	60.20	**66.68**	59.82	62.33
HA	46.69	55.14	53.08	67.00	66.14	62.83	**73.20**	66.25	65.44	65.06
SP	44.80	53.35	51.94	64.20	**71.58**	58.93	66.70	68.24	67.46	70.27
HC	46.30	48.22	53.58	50.20	62.16	47.67	64.90	**65.21**	60.05	64.29

Table 4. Results comparison with the state-of-the-art methods on HSRC2016 (%).

Methods	mAP (%)	Inference Speed (FPS)
R2CNN [4]	73.07	3
RC1 & RC2 [34]	75.70	1
RRPN [12]	79.08	4.8
R2PN [13]	79.60	1.3
Tian et al. * [35]	80.80	4
RetinaNet [17]	82.89	16
RoI-Trans. [5]	86.20	9
R3Det [17]	89.14	6
SCRDet [14]	89.27	0.8
SAFDet (proposed)	**89.38**	11

* Result is evaluated by DOTA evaluation metrics.

4.3.2. HRSC2016

Table 4 illustrates the results in comparison to the state-of-the-art methods on the HRSC2016 dataset. As seen, the proposed method reached 89.38% mAP without adopting any image pyramid scheme—i.e., the longest side of image was resized to 800 as used in [5,12,17]—thereby outperforming all other methods. The inference speed with respect to each detector is also listed in Table 4.

For the result comparison in terms of mAP, the proposed method achieved 89.38%, outperforming SCRDet by 0.11%. We deduced that the feature extraction modules proposed in SCRdet were over complicated for HRSC2016. For the result comparison in terms of inference speed, however, the proposed SAFDet was about 14 times faster than SCRDet (11 FPS vs. 0.8 FPS), which further validates that our SAFDet method has apparent advantages in balancing between the accuracy and the computational cost.

The network architecture of Tian et al. [35] is similar to that of the proposed SAFDet. However, its inference speed is 4 fps, which is 64% slower than our proposed SAFDet. This was mainly due to the utilization of a rotational anchor on the RPN, which has low inference efficiency compared with the horizontal anchor utilized in the SAFDet.

5. Conclusions

In this paper, we proposed a semi-anchor-free detector (SAFDet) for effective detection of oriented objects in aerial images. Two new modules are introduced to tackle the detection difficulties of low-resolution, noisy, large aspect ratio and freely-oriented objects. A rotational anchor-free branch is introduced to assist the HBB prediction on RPN, with negligible extra computational cost during the training. Similarly to the attention module, CPM is proposed to suppress the background information and enhance the foreground information. However, CPM is only implemented for training. By adopting those two modules, the proposed model improves the mAP by about 3% without bringing any additional computational cost to inference. When compared with the state-of-the-art detectors, the proposed method achieves effective results on the challenging DOTA and HRSC2016 datasets, without lowering its high efficiency.

As mentioned above, the existing anchor-free matching schemes were designed for HBB regression, which inspired us to propose a new matching scheme for OBB regression in the future. Furthermore, we will validate the performance on other datasets, such as VEDAI [36], a dataset containing aerial and satellite images collected with a fixed spatial resolution, as well as applying new models such as saliency detection [37], improved feature extraction [38,39] and deep learning [40] for more effective object detection from remote sensing images.

Author Contributions: Implementation and drafting the paper, Z.F.; conceptual design, intensive review/editing and funding, J.R.; data collection and preprocessing, H.S.; writing—review and editing, S.M.; conceptual design and writing—review and editing, J.H.; formal analysis and writing—review and editing, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially funded by Dazhi Scholarship of Guangdong Polytechnic Normal University (991620475), National Natural Science Foundation of China (62072122), and Education Dept. of Guangdong Province (2019KSYS009).

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338.
2. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
3. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415.
4. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational region cnn for orientation robust scene text detection. *arXiv* **2017**, arXiv:1706.09579.
5. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
6. Liao, M.; Shi, B.; Bai, X. Textboxes++: A single-shot oriented scene text detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690.
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
8. Girshick, R. Fast R-cnn. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Washington, DC, USA, 7–13 December 2015; pp. 1440–1448.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2017**, *39*, 1137–1149.
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-cnn. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
11. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078.
12. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122.
13. Zhang, Z.; Guo, W.; Zhu, S.; Yu, W. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1745–1749.
14. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 8232–8241.
15. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Körner, M.; Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. In Proceedings of the 2018 Asian Conference on Computer Vision (ACCV), Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 150–165.
16. Wang, J.; Yuan, Y.; Yu, G. Face attention network: An effective face detector for the occluded faces. *arXiv* **2017**, arXiv:1711.07246.
17. Yang, X.; Liu, Q.; Yan, J.; Li, A. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *arXiv* **2019**, arXiv:1908.05612.
18. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.

19. Mathias, M.; Benenson, R.; Pedersoli, M.; Van Gool, L. Face detection without bells and whistles. In Proceedings of the 2014 European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 720–735.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
25. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the 2019 IEEE ICCV, Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
26. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. *arXiv* **2019**, arXiv:1904.01355.
27. Liu, W.; Liao, S.; Ren, W.; Hu, W.; Yu, Y. High-level semantic feature detection: A new perspective for pedestrian detection. In Proceedings of the 2019 IEEE Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5187–5196.
28. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Bradski, G.; Kaehler, A. The OpenCV library. *Dr. Dobb's J. Softw. Tools* **2000**, *120*, 122–125.
31. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.
32. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R. RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sens.* **2020**, *12*, 389.
33. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308.
34. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the 2017 International Conference on Pattern Recognition Applications and Methods (ICPRAM), Porto, Portugal, 24–26 February 2017; Volume 2, pp. 324–331.
35. Tian, T.; Pan, Z.; Tan, X.; Chu, Z. Arbitrary-Oriented Inshore Ship Detection based on Multi-Scale Feature Fusion and Contextual Pooling on Rotation Region Proposals. *Remote Sens.* **2020**, *12*, 339.
36. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203.
37. Yan, Y.; Ren, J.; Sun, G.; Zhao, H.; Han, J.; Li, X.; Marshall, S.; Zhan J. Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement. *Pattern Recognit.* **2018**, *79*, 65–78.
38. Zabalza, J.; Ren, J.; Zheng, J.; Zhao, H.; Qing, C.; Yang, Z.; Du, P.; Marshall, S. Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. *Neurocomputing* **2016**, *185*, 1–10.

39. Zabalza, J.; Ren, J.; Yang, M.; Zhang, Y.; Wang, J.; Marshall, S; Han, J. Novel folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2014**, *93*, 112–122.
40. Zabalza, J.; Ren, J.; Zheng, J.; Han, J.; Zhao, H.; Li, S.; Marshall, S. Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4418–4433.